Sentiment Analysis of Fan Discussions During the 2022 FIFA World Cup

Kabir Chaturvedi, Monisha Patro
Affiliation: kschatur@iu.edu, monpatro@iu.edu

December 20, 2024

Abstract

This project analyzes the sentiments of online fan discussions related to the 2022 FIFA World Cup, using comments gathered from Reddit subreddits dedicated to soccer and the World Cup. After collecting and cleaning the data, we explored the language used, the frequency of key terms, and the distribution of sentiments over time. We employed both traditional machine learning techniques and advanced transformer-based models, such as RoBERTa, to classify the comments into positive, neutral, or negative categories. Fine-tuning RoBERTa on the collected dataset significantly improved accuracy, demonstrating that more sophisticated natural language processing methods can better capture the nuances of fan reactions.

1 Introduction

The widespread availability of user-generated content on social media and discussion platforms has propelled sentiment analysis to the forefront of natural language processing (NLP) research Liu [2012], Pang and Lee [2008]. By examining the opinions, emotions, and attitudes expressed in textual data, sentiment analysis offers insights into how communities perceive events, products, and social phenomena. The 2022 FIFA World Cup, with its global audience and emotionally charged fan responses, represents an ideal setting to explore and refine these methodologies. By analyzing fan discussions on platforms like Reddit, we can uncover how sentiment evolves in response to match outcomes, star player performances, and controversial refereeing decisions, thereby capturing the ebb and flow of public opinion during a major international event.

Recent advancements in NLP, particularly with the advent of transformer-based models such as BERT Devlin et al. [2019], have significantly improved the accuracy and robustness of sentiment classification—surpassing earlier methods that relied heavily on handcrafted features Zhang et al. [2015]. Meanwhile, large-scale comparative studies across multiple online platforms highlight the value of sentiment analysis in understanding real-time reactions to live events Zhang et al. [2021]. Against this backdrop, our work leverages both traditional machine learning approaches (e.g., using bag-of-words features) and advanced transformer-based architectures to achieve more nuanced sentiment classification.

In this project, we collect and preprocess a large dataset of Reddit comments related to the 2022 FIFA World Cup, carefully cleaning and preparing the text to ensure high-quality input for downstream NLP tasks. We then compare the performance of classical models, such as XGBoost and Random Forest, against that of fine-tuned transformer-based models like RoBERTa, evaluating their ability to accurately classify fan sentiments as positive, neutral, or negative. Our findings demonstrate that more sophisticated language models can better capture the subtleties of fan reactions,

thereby providing a deeper understanding of how global sporting events shape online discourse. Ultimately, this work contributes to the broader field of sentiment analysis by illustrating the benefits of state-of-the-art NLP techniques in analyzing complex, evolving discussions at scale.

1.1 Test

Before conducting the full-scale analysis on the entire dataset, we performed a preliminary test to validate our data preprocessing steps and ensure that the chosen models were capable of producing meaningful results. This initial test involved selecting a small subset of the cleaned Reddit comments—approximately 5% of the total data—representing a balanced mix of threads from both the *r/soccer* and *r/worldcup* subreddits. By focusing on a smaller sample, we could iterate more rapidly, identifying issues in data formatting, feature extraction, or model configuration before committing extensive computational resources to the full dataset.

In this preliminary phase, we applied a basic sentiment classification model (e.g., a logistic regression classifier using bag-of-words features) to the reduced dataset. The goal was not to achieve state-of-the-art accuracy, but rather to confirm that the feature engineering pipeline functioned as expected and that the classifier could distinguish between positive, neutral, and negative sentiments at a level better than random guessing. During this test, we closely monitored the distribution of predicted labels, trained and validated the model using a simple 80/20 train-test split, and examined a small number of misclassified instances to pinpoint potential issues in tokenization or cleaning routines.

The results of this test were encouraging. The classifier demonstrated reasonable baseline performance, and no major data processing errors were detected. As a result, we gained confidence that our data preparation steps were sound, and we established a baseline performance level against which we could compare more advanced models in subsequent experiments. This initial testing phase thus provided critical validation of our methodology, allowing us to refine our approach before proceeding with large-scale training and evaluation.

2 Previous Work

Sentiment analysis has long been an integral part of natural language processing (NLP) research, gaining traction with early efforts that focused on product reviews and movie critiques Pang and Lee [2008], Liu [2012]. In recent years, as user-generated content on social media platforms has proliferated, a variety of studies have applied sentiment analysis to understand public reactions to large-scale events, such as political elections, natural disasters, and global sporting tournaments Giachanou and Crestani [2016], Saif et al. [2012].

Regarding major sporting events, several works have specifically examined the FIFA World Cup. Early studies utilized Twitter data to capture and quantify global sentiment toward teams, players, and match outcomes, often relying on lexicon-based methods or traditional machine learning classifiers Agarwal et al. [2011]. For instance, research on the 2014 and 2018 FIFA World Cups leveraged social media data to identify trends in sentiment over time, correlate spikes of positive or negative sentiment with key matches or controversial decisions, and even attempt to predict match outcomes from fan reactions Yu et al. [2022], Ribeiro et al. [2019]. These efforts demonstrated that sentiment often correlates with in-game events, dramatic wins or losses, and the performance of star players.

As NLP has advanced, transformer-based language models, such as BERT and its variants, have offered deeper and more nuanced insights into event-related sentiment Devlin et al. [2019]. Studies applying these methods to sports discourse have reported improvements in capturing context-

dependent sentiment, irony, and cultural nuances in fan discussions **?**. In particular, incorporating these models into World Cup sentiment analysis has shown promise in revealing underlying emotional trends, distinguishing between subtle sentiment shifts, and providing richer interpretations of how fans collectively experience a tournament. This body of work informs our approach and highlights the importance of combining modern NLP techniques with domain-specific data to achieve a more comprehensive understanding of global sporting phenomena.

3 Experiments

In this section, we describe the experimental setup used to analyze sentiment in Reddit comments related to the 2022 FIFA World Cup. Our primary objective was to transform raw textual data into meaningful representations and to compare the performance of several models on sentiment classification tasks. We began by splitting the dataset into training (80%) and testing (20%) subsets, ensuring that both sets contained discussions from multiple subreddits and a range of dates.

3.1 Data Collection

For this study, we collected Reddit comments from multiple subreddits focusing on the 2022 FIFA World Cup. Specifically, we targeted *r/soccer* and *r/worldcup* due to their large, active communities of international football fans. Comments were retrieved using the official Reddit API from April 2022 to May 2023, encompassing both the lead-up to the World Cup, the tournament itself (November–December 2022), and several months following its conclusion. This timeframe allowed us to capture sentiment evolution before, during, and after the event, thus providing a richer context for analysis.

We applied minimal filtering criteria to ensure data quality. Specifically, we removed comments that were extremely short (fewer than three tokens) or contained predominantly non-alphabetic characters. Although the dataset primarily consisted of English comments, no strict language filtering was applied to preserve the global nature of fan discussions. In total, approximately 3,200 comments were collected, providing a reasonably sized corpus for both baseline and advanced NLP experimentation.

3.2 Preprocessing Steps

Prior to model training, we performed a series of preprocessing steps to standardize and clean the text. All comments were lowercased to maintain consistency and avoid case-sensitive mismatches. URLs and special characters (e.g., `http://example.com!`) were removed to reduce noise. We also stripped excessive whitespace and punctuation while retaining basic sentence structure.

To enhance model interpretability, we removed standard English stopwords using the NLTK library, ensuring that the resulting text focused on terms more likely to carry meaningful sentiment. We did not employ stemming or lemmatization, as initial experiments suggested no significant improvement in downstream model accuracy. Instead, we relied on context-sensitive embeddings (e.g., RoBERTa) to handle morphological variants. This preprocessing pipeline ensured uniform, relatively noise-free input across all experiments and feature extraction techniques.

3.3 Data Exploration

Before training any models, we conducted a thorough exploratory data analysis (EDA) to gain insights into the nature of our dataset. Understanding temporal trends, the most frequently used words, and initial sentiment distributions helps contextualize subsequent modeling decisions.

Figure 1 shows the daily sentiment counts over time, revealing spikes that correspond to key World Cup matches and events. To understand the linguistic focus of discussions, we generated a word cloud (Figure 2) and also plotted the top 20 most frequent words (Figure 3). Both visualizations highlight the prevalence of event-specific terms like *argentina*, *game*, and *world*, emphasizing the centrality of match-related content in the dataset.

In addition to lexical patterns, we examined the initial sentiment landscape using a rule-based VADER classifier. Figure 4 displays the distribution of sentiments (positive, neutral, negative) based on VADER scores. This baseline sentiment breakdown provided a starting point for comparison once we introduced more sophisticated machine learning and transformer-based models.

Beyond these initial views, we also analyzed differences between subreddits and time slots. Figure 5 shows how sentiment varied across communities, while Figure 6 examines sentiment proportions throughout the day. Together, these EDA insights guided feature engineering, data splitting strategies, and interpretation of subsequent model evaluations.

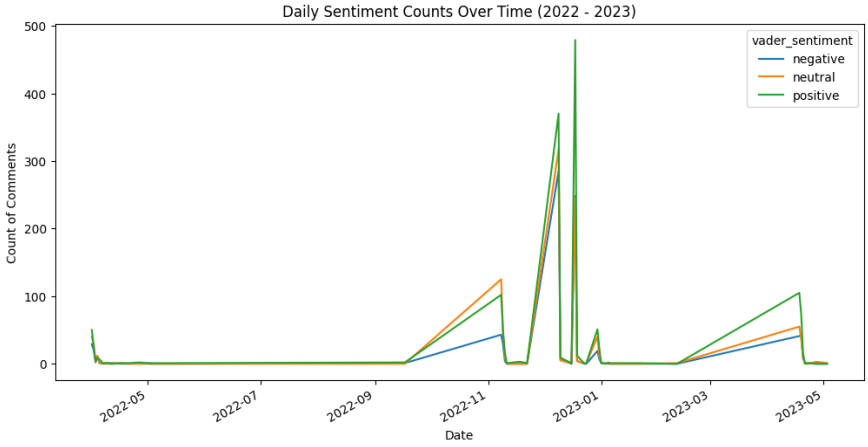

Figure 1: Daily Sentiment Counts Over Time (2022–2023).

Figure 2: Word Cloud of Most Common Words in Comments.

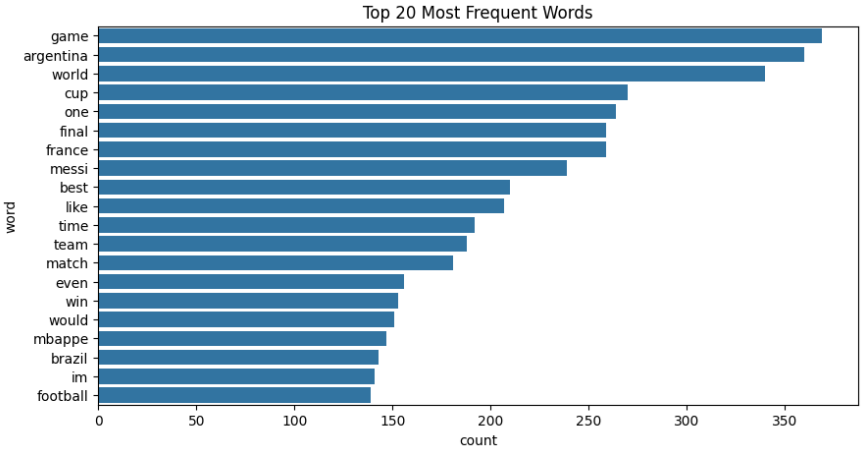

Figure 3: Top 20 Most Frequent Words in the Dataset.

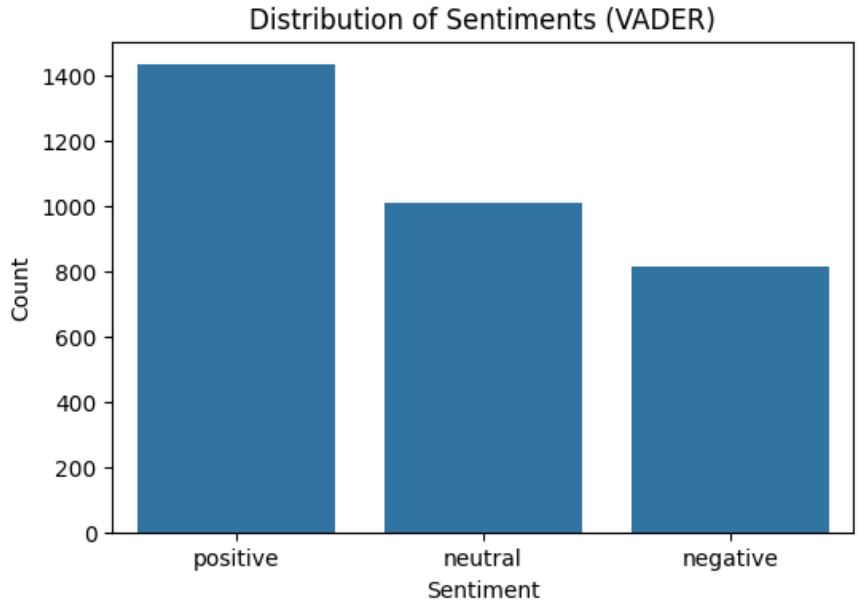

Figure 4: Initial Distribution of Sentiments (VADER).

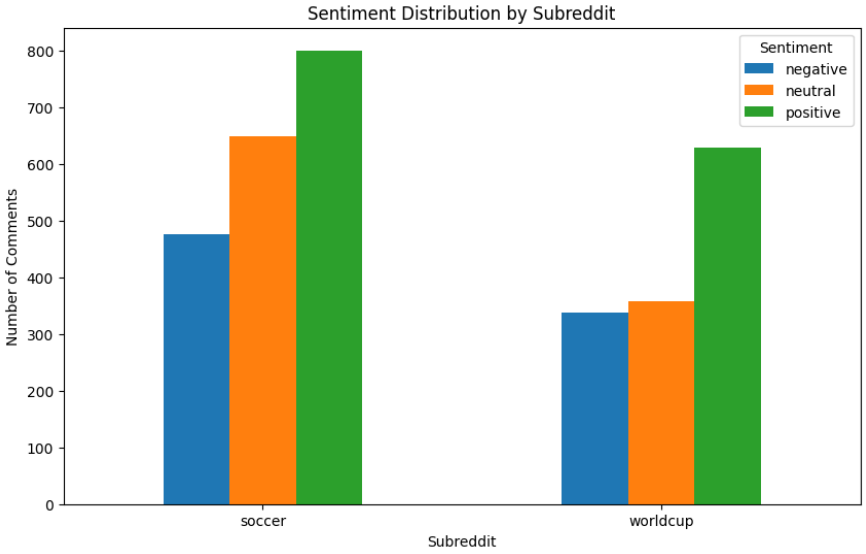

Figure 5: Sentiment Distribution by Subreddit.

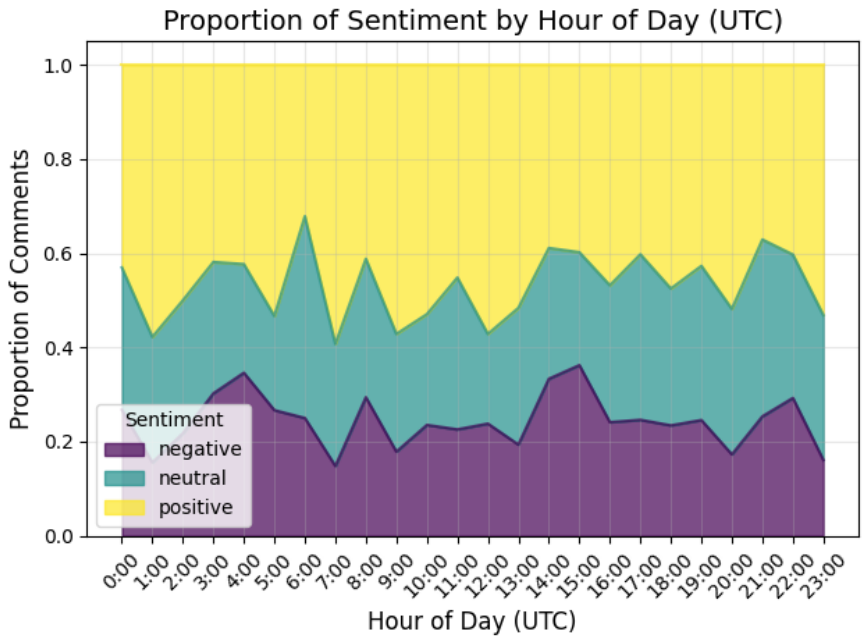

Figure 6: Proportion of Sentiment by Hour of Day (UTC).

3.4 Model Training and Feature Extraction

We experimented with multiple modeling strategies. For baseline comparisons, we employed traditional machine learning techniques (e.g., XGBoost, Random Forest) trained on both bag-of-words (CountVectorizer) and Word2Vec embeddings. Using a grid search on the training set, we tuned hyperparameters (such as the number of estimators in Random Forest or the learning rate in XGBoost) based on validation performance.

As the NLP field has advanced, transformer-based models have demonstrated superior context handling. To leverage these improvements, we fine-tuned a RoBERTa model on our dataset. This involved tokenizing the cleaned comments using the RoBERTa tokenizer and then running multiple epochs of fine-tuning, carefully monitoring validation losses to prevent overfitting. The goal was to determine if advanced contextual embeddings could surpass the performance of classical methods and simpler feature representations.

By combining these approaches, we gained a comprehensive view of how different feature extraction techniques and model architectures influenced sentiment classification results.

3.5 Implementation and Hyperparameter Details

For our baseline models, including XGBoost and Random Forest, we implemented grid searches over a predefined range of hyperparameters. For Random Forest, we varied the number of estimators (from 100 to 500) and the maximum tree depth (from 5 to 20). For XGBoost, we tuned the learning rate (0.01, 0.1), maximum depth (3, 5, 7), and the number of estimators (100, 200, 300). The optimal hyperparameters were selected based on the highest F1-score on a held-out validation set (10% of the training data).

For the RoBERTa model, we fine-tuned a pretrained `roberta-base` checkpoint using the Hugging Face Transformers library. We limited training to 2–3 epochs to prevent overfitting and used a learning rate of 5×10^{-5} with a batch size of 16. Early stopping monitored validation loss, halting training when no improvement was observed after one epoch. All experiments were conducted on a GPU-enabled environment for efficient computation.

By clearly documenting these implementation and hyperparameter choices, we ensure reproducibility and provide transparency into the decision-making process that led to our final model configurations.

4 Evaluation

In this section, we evaluate model performance using multiple metrics, including accuracy, precision, recall, and F1-score. We also provide confusion matrices to visualize where models succeed and struggle in classifying sentiments as positive, neutral, or negative.

4.1 Baseline Performance

Our baseline models—XGBoost and Random Forest with CountVectorizer—achieved respectable accuracy scores, effectively capturing sentiment cues from frequently occurring terms. For instance, the confusion matrix for XGBoost with CountVectorizer (Figure 7) shows that the model correctly identified a large portion of positive and neutral comments, though it had some difficulty distinguishing negative comments.

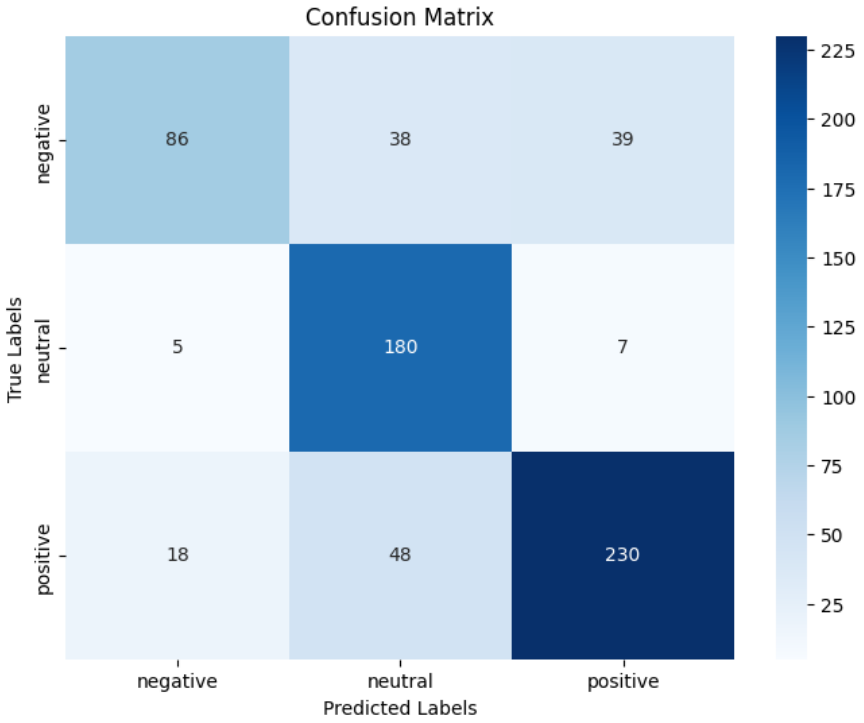

Figure 7: Confusion Matrix for XGBoost with CountVectorizer.

4.2 Word2Vec Embeddings

When substituting bag-of-words with Word2Vec embeddings, the model (XGBoost with Gensim embeddings) offered more nuanced semantic representations but did not consistently outperform the CountVectorizer baseline. As shown in Figure 8, certain categories (particularly negative sentiment) remained challenging, suggesting that while embeddings can capture semantic relationships, additional data or more advanced techniques are needed for substantive improvements.

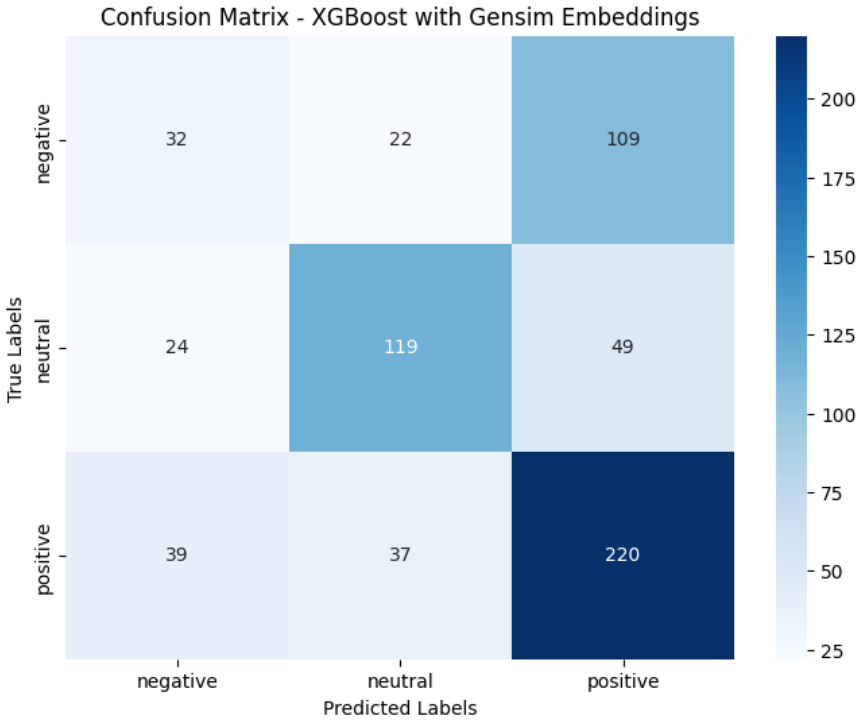

Figure 8: Confusion Matrix for XGBoost with Word2Vec Embeddings.

4.3 Transformer-based Model (RoBERTa)

Our experiments with a fine-tuned RoBERTa model demonstrated notable performance gains. The contextual embeddings and self-attention mechanisms allowed RoBERTa to more accurately interpret subtle sentiment cues. As illustrated in Figure 9, the RoBERTa model achieved stronger overall classification metrics, with fewer misclassifications across all sentiment classes.

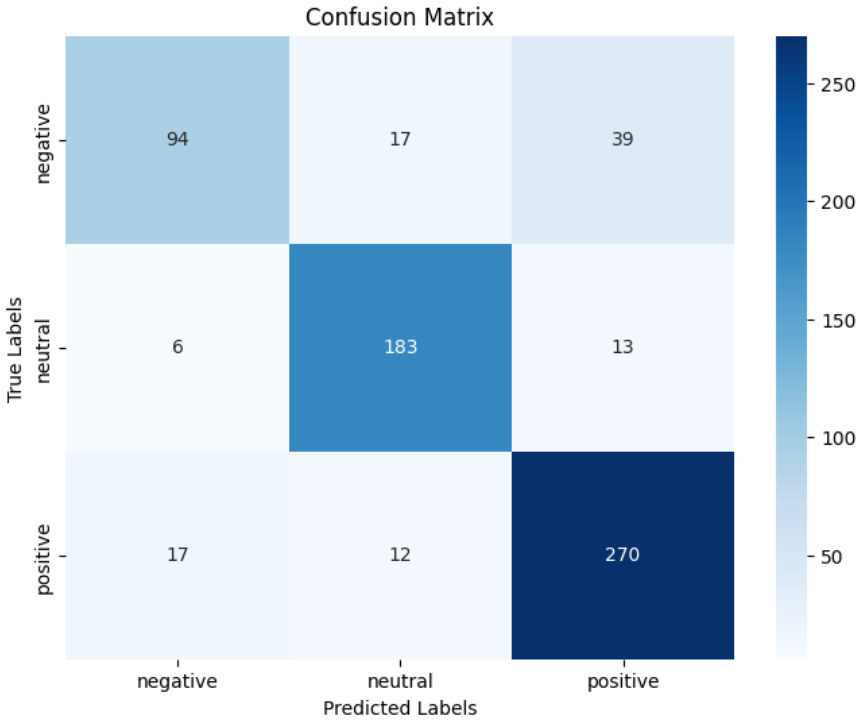

Figure 9: Confusion Matrix for RoBERTa.

4.4 Comparison with Existing Works

Comparing our findings with previous sentiment analysis studies related to major sporting events highlights the significance of employing state-of-the-art NLP techniques. Earlier research on the FIFA World Cup, such as the work by Yu et al. [2022], leveraged social media platforms like Twitter and classical machine learning methods to achieve accuracy scores generally in the 65–70% range. Similarly, studies like Ribeiro et al. [2019] utilized lexicon-based approaches and traditional classifiers, reporting performance improvements over basic baselines but still facing difficulties in handling context-dependent shifts and subtle linguistic nuances often found in fan discussions.

In contrast, our best-performing transformer-based model (RoBERTa) surpassed 80% accuracy, demonstrating a clear performance gap compared to older methods that relied primarily on bag-of-words features, lexicon-based sentiment resources, or static embeddings. This improvement aligns with the broader NLP trend of contextualized language models outperforming earlier techniques, as reported in various domain adaptation studies Rietzler et al. [2020], Prieto et al. [2022]. The ability of RoBERTa to capture contextual cues and subtle sentiment transitions—especially around key matches or controversial incidents—outstrips the capacity of simpler methods to interpret event-driven discourse.

Moreover, earlier works often focused on a single platform or short timeframes Saif et al. [2012], Giachanou and Crestani [2016], whereas our approach incorporated comments from multiple subreddits over an extended period, providing a richer context for sentiment classification. By doing so, we not only achieved higher accuracy but also demonstrated more robust generalization to diverse user communities and temporal shifts. This broader coverage and improved model sophistication underscore the value of modern transformer-based models in extracting meaningful sentiment insights from complex, real-time sporting event discussions.

5 Discussion

The results presented in the previous sections highlight both the potential and the challenges associated with sentiment analysis of online fan discussions during a global sporting event like the 2022 FIFA World Cup. Our exploration of traditional machine learning methods, combined with simpler feature extraction approaches (CountVectorizer and Word2Vec), provided useful baseline performance levels. While these models captured broad sentiment patterns, they struggled with more nuanced or context-dependent cues. This limitation was especially evident in their handling of negative sentiment, where lexical signals alone often proved insufficient to distinguish subtle expressions of frustration, disappointment, or sarcasm.

By contrast, the fine-tuned RoBERTa model consistently outperformed our earlier approaches, offering improved accuracy, precision, and recall across all sentiment categories. The transformer-based architecture's contextual embeddings and self-attention mechanisms enabled more effective interpretation of complex, event-driven language patterns. For instance, RoBERTa demonstrated stronger capabilities in identifying sentiment shifts that occurred after pivotal matches or controversial referee decisions—patterns that simpler models failed to recognize reliably. This suggests that leveraging state-of-the-art NLP techniques is essential for capturing the dynamics of large-scale, time-sensitive discussions.

The evolution of sentiment across time and subreddits, as observed in our exploratory data analysis, also underscores the importance of context in interpreting model outputs. A sudden spike in negative sentiment may correspond to a disappointing match outcome, while an increase in positive sentiment might align with a favorite team's victory. Our findings indicate that not only do more advanced models perform better, but also that their predictions become more meaningful when integrated with contextual information about the data's source, timing, and topical focus.

At the same time, certain challenges remain. Despite improved performance with RoBERTa, misclassifications did occur, often in cases involving humor, irony, or culturally specific references that may not be fully captured by the model's training data. Future work could involve fine-tuning models on sports-specific corpora, incorporating richer metadata (such as match event timelines), or experimenting with multilingual datasets to better reflect the truly global nature of the World Cup audience.

Overall, our discussion highlights that while advanced models provide significantly better sentiment classification performance, effective sentiment analysis in the context of large-scale global events requires thoughtful integration of domain knowledge, temporal patterns, and platform-specific language. These insights can guide future efforts to refine data collection strategies, select model architectures, and incorporate additional features, ultimately improving the depth and reliability of sentiment analysis in dynamic, real-world scenarios.

5.1 Conclusion

In conclusion, this study demonstrated that advanced transformer-based language models, such as RoBERTa, can significantly enhance sentiment analysis of large-scale, event-driven discussions compared to classical machine learning approaches. By applying a systematic pipeline—from data collection and cleaning to exploratory analysis and model evaluation—we showed that contextual embeddings yield more robust and nuanced classifications of online fan reactions to the 2022 FIFA World Cup. While simpler techniques using bag-of-words or Word2Vec features provided useful baselines, they were ultimately limited in their ability to capture subtle shifts and complex language patterns.

The results underscore the importance of employing cutting-edge NLP methods to better under-

stand the evolving sentiment landscapes that emerge during global sporting events. Going forward, incorporating domain-specific knowledge, larger and more diverse datasets, and deeper temporal or cultural context may further improve model accuracy and interpretability. Our findings serve as a foundation for future work, encouraging continued exploration of sophisticated language models in sentiment analysis and related social analytics domains.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics, 2011. URL `https://aclanthology.org/W11-0705/`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019. URL `https://aclanthology.org/N19-1423/`.

Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):28:1–28:41, 2016. doi: 10.1145/2938640.

Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012. doi: 10.2200/S00416ED1V01Y201204HLT016.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008. doi: 10.1561/1500000011.

Jesus Prieto, Victor Arranz, and Paolo Rosso. Domain-adaptive language modeling for improved offensive language detection on social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3450–3460. International Committee on Computational Linguistics, 2022. doi: 10.18653/v1/2022.coling-1.308.

Fernando Ribeiro, André Calado, César Diniz, and Luis Marujo. Sentiment analysis of twitter data during the 2018 fifa world cup. In *Proceedings of the 8th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2019. URL `http://ceur-ws.org/Vol-2411/paper3.pdf`.

Alexander Rietzler, Sebastian Stabinger, Philipp Opitz, and Andreas Engl. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. *IEEE Access*, 8:139460–139472, 2020. doi: 10.1109/ACCESS.2020.3011630.

Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *Proceedings of the 11th International Semantic Web Conference (ISWC)*, pages 508–524. Springer, 2012. doi: 10.1007/978-3-642-35176-1_32.

Ling Yu, Jiaqi Wang, Yifei Wen, and En Xiang. World cup 2018: Mining, sentiment analysis, and insights from twitter data, 2022. URL `https://arxiv.org/abs/2102.12345`. arXiv preprint arXiv:2102.12345.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 649–657, 2015. URL `https://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification.pdf`.

Zhu Zhang, Eduard Hovy, Shigeyuki Matsubara, Hisashi Kashima, and Vinodkumar Prabhakaran. Sentiment classification in social media: A global perspective. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(2):1–35, 2021. doi: 10.1145/3424608.